The Pharmacogenomics Journal (2003), 1–4

www.nature.com/tpj

Class struggle: expression profiling and categorizing cancer

PA Covitz

National Cancer Institute Center for Bioinformatics, 6116 Executive Boulevard, Rockville, MD, USA

The Pharmacogenomics Journal advance online publication, 19 August 2003; doi:10.1038/sj.tpj.6500187

INTRODUCTION

The classification of disease by characterization of the morphologic features of affected tissue is a conventional approach to establish a diagnosis. Upon inspection of a tissue sample under light microscopy, a pathologist describes the morphologic findings applying commonly used diagnostic terms. The NCI Thesaurus, which captures terminology in use by the National Cancer Institute (NCI), lists over 6000 neoplastic disease concepts with approximately 24 000 preferred names and synonyms.[1] The NCI Metathesaurus, an even broader system that maps synonymy across many biomedical vocabulary sources, suggests that as many as 7000 concepts with 42 000 preferred names and synonyms for neoplastic disorders may be in use worldwide.[2] While there is no doubt great diversity in the manifestation of neoplasms in the human population, this pool of terminology likely exceeds the necessary descriptive and explanatory complexity that might be needed to properly distinguish scientifically and clinically distinct diseases.

There is thus an increasing awareness of the need to consolidate and standardize disease terminology, and modernize the basis by which distinct cancer types are defined. The International Classification of Disease for Oncology (ICD-O) from World Health Organization (WHO) is perhaps the most comprehensive effort to organize descriptions of cancer into coded categories.[3] The ICD-O organizes classifications into two major axes: morphology and topography. The third edition, ICD-O-3, includes a substantial update to the classification of hematologic malignancies. In addition to morphological characteristics, the criteria used to classify the hematologic disorders include cytogenetic abnormalities, immunophenotypic profiles, and clinical presentation. The ICD-O is intended for coded summarization of disease incidence reporting rather than for clinical practice. A volume that gives greater detail on the histopathologic, immunologic, and cytogenetic characteristics of each disease class is also available.[4]

The inclusion of cytogenetic features into the WHO classification reflects the importance of capturing underlying molecular phenomenon that are associated with particular disease subtypes, in some cases warranting an entire classification subbranch (Table 1). At present, cytogenetic aberrations are the principal type of genetic observation used in classifying some subcategories of specific hematologic malignancies (acute myeloid leukemia and myelodysplastic syndromes). The availability of data correlating chromosomal rearrangements and gene fusion events with specific diseases, as well as the standardization of how such translocations are described, has made it possible to include this type of genetic information in a formal classification.

Chromosomal rearrangements and their associated gene fusions represent one of several types of genetic abnormalities that can contribute to the etiology of cancer. Aberrations such as hyperploidy, gene amplification, and sequence mutations can also be involved or correlated with the establishment or metastasis of malignancy. The immediate phenotypic consequences of any of these genomic DNA lesions can be measured in terms of altered gene expression patterns. It is therefore appropriate to consider how, along with morphologic, immunophenotypic, and cytogenetic characteristics, canonical gene expression patterns might be used as classifiers in standardized descriptions of cancer. Recent gene expression profiling studies of hematologic and other types of malignancies illustrate the progress and challenges in pursuing this end.

Diffuse Large B-cell Lymphoma (DLBCL)

In the present WHO classification, DLBCL is not subgrouped according to any distinguishing molecular characteristics. Patients diagnosed with DLBCL vary widely in their clinical courses and outcomes, suggesting there are important differences in the underlying disease etiology and responses to treatment across this population.[5,6]

Starting with the sequence and cDNA library resources available from the Cancer Genome Anatomy Project, a team of collaborators at several institutions developed a so-called lymphochip cDNA microarray to probe molecular profiles of gene expression in lymphocytic malignancies.[7] They made use of the lymphochip in a comparative analysis of 96 normal and malignant lymphocyte samples.[8] After unsupervised hierarchical clustering of the resulting data, the authors qualitatively characterized several of the clusters as 'signatures' and labeled them according to a biological feature shared by many of the genes within the cluster. Genes in the 'Germinal Center B Cell' signature

Table 1 Excerpt of WHO classification of hematopoietic and lymphoid neoplasms*

ICD-O	*Classification*
	Acute myeloid leukemias (AMLS)
	AMls with recurrent cytogenetic translocations
9896/3	AML with {t(8;21)(q22;q22)}, {AML1(CBF-alpha)/ETO}
9866/3	Acute promyelocytic leukemia {AML with t(15;17)(q22;q11-12} and variants, {PML/RAR-alpha}
9871/3	AML with abnormal bone marrow eosinophils {inv(16)(p13q22)} or {t(16;16)(p13;q22)}, {CBFb/MYH11}
9897/3	AML with 11q23 abnormalities {MLL}
9895/3	AML with multilineage dysplasia
9895/3	With prior myelodysplastic syndrome
9895/3	Without prior myelodysplastic syndrome
9920/3	AML and myelodysplastic syndromes, therapy-related
9920/3	Alkylating agent-related
9920/3	Epipodophyllotoxin-related (some may be lymphoid)
9920/3	Other types

*Source: http://training.seer.cancer.gov/module_coding_primary/table_who_class_hemo_1.html.

were chosen to serve as a basis for further DLBCL subtyping, based on the view that DLBCL may descend from different B-cell developmental stages. Indeed, the signatures from each of the diseased patient samples were found to segregate into one of two classes: one similar to normal germinal center B cells, the other similar to activated B cells.

Overall, patients with activated B-like DLBCL had significantly lower survival rates than those with germinal B-like DLBCL. This finding suggests that gene expression profiling can be used to subclassify DLBCL into at least two diseases. A more recent lymphochip study with a larger number of samples confirmed this finding, and further demonstrated that expression profiling could be used to develop a prognostic indicator that is more accurate than conventional indices.[9] However, the results from the analysis of commercial oligonucleotide array data from a different cohort of samples using supervised machine learning methods did not find correlation between B-cell developmental stage signature and clinical outcome.[10] The supervised approach defined clinical outcome as the basis for classification at the outset, and identified genes whose expression signatures provide high predictive power. The differing results from these studies stem from the different array technology, array design, analytical approach, and patient cohorts that were involved, and illustrate the difficulty in interpreting and comparing microarray-based findings from different sources. Readers are directed to a recent article by D Slonim for a thorough treatment of microarray analysis approaches, including the distinction between finding patterns *vs* finding classifiers.[11] Chuaqui *et al*[12] provide a cogent discussion of the sources of uncertainty, variation and error in microarray experimental design and analysis.

Acute Lymphoblastic Leukemia (ALL)

Cytogenetic and immunophenotyping analyses have established two broad categories of ALL.[13] B-cell lineage ALL (B-ALL) can be further subclassified in part by the chromosomal translocations that result in particular gene fusions or by hyperploidy. T-cell lineage ALL (T-ALL) is also correlated with cytogenetic abnormalities, although these appear less frequently than in the B-ALL population.

A relatively broad study probed 327 bone marrow samples with commercial oligonucleotide arrays and found that seven distinct leukemia subtypes could be defined through unsupervised hierarchical clustering of the data.[14] One of these subtypes corresponded to T-ALL; five corresponded B-ALL subtypes previously classified according to cytogenetic rearrangements or hyperploidy. These results are a remarkable independent confirmation of significance of these classes as biologically distinct subtypes. The seventh subtype emerged from a group of patients who lacked any consistent chromosomal rearrangement and had varying degrees of ploidy ranging from normal to hyperdiploid. In total, 20% of the cases did not cluster into any of the seven leukemia subtypes. These finding suggest that at least one and perhaps several additional distinct disease subclasses are prevalent in the ALL patient population.

Support vector machine analysis of a training set of ALL gene expression signatures resulted in the identification of representative genes within the clusters that can be used as predictive classifiers.[14] The number of genes needed to classify a given subtype ranged from one to 20. When these classifiers were used to analyze a test set of signatures that were not part of the training set, they yielded accuracy, sensitivity, and specificity values of 93–100% for nearly all three measurements of each ALL subtype, a remarkable level of predictive power.

Another expression profiling study focusing specifically on T-ALL defined several subtypes that correlated with known oncogene activation and, further, identified a newly suspected oncogene.[15] The microarray results were confirmed with RT-PCR, and, as with the other studies, demonstrated that expression signatures could be used as prognostic indicators.

Other Cancers

Microarray technology has been applied to several other types of cancer by scientists attempting to discern molecular subclasses. A study analyzing 65 biopsy specimens from 42 breast cancer patients found a variety of patterns across tumors from different patients, making it difficult to clearly articulate a diagnostic portrait of cancer subtypes. Nonetheless, the analysis was able to provide a preliminary indication that distinct molecular classes exist, and could perhaps be better distinguished and characterized with larger data sets.[16]

Another study, reported results from an attempt to use microarrays to classify lung tumors.[17] The researchers collected data from 41 adenocarcinomas (ACs), 16 squamous cell carcinomas (SCCs), five large cell lung cancers (LCLCs), and five small cell lung cancers (SCLCs) as determined by a pathologist using light microscopy. Hierarchical cluster analysis showed strong correlation between particular clusters and the morphological classification of the original samples, an appealing validation of the morphology-based classification system in use for lung cancer. The analysis further revealed that the AC class could be divided into three subclasses, but the clinical significance of the subdivision was not entirely clear. As with the breast cancer study, the authors suggest a larger study with more samples would provide more compelling evidence for clinically relevant lung tumor subclasses based on expression profile data.

Classification Using Expression Signatures

The examples of DLBCL and ALL cited above illustrate the power and potential for using gene expression profiling in classifying specific disease subtypes and predicting clinical outcome. The case for subclassification of solid tumors is less compelling at present, but as research data accumulate, an increasing number of small sets of genes that can be used for unambiguous disease classification will likely be identified.

A number of important challenges remain before such techniques can be formally adopted. A major issue is the adolescent state of microarray technology. Heterogeneity of results across the various platforms,[8,10] and a tremendous diversity of algorithmic approaches one can take to the data,[11] makes it difficult to precisely define what 'microarray analysis' consists of. Further, microarray designs are changing as data from the human genome sequence and associated exon mapping studies make their way into the literature and public databases.[18,19] Individual investigators can make choices about what technological approaches and array designs to use for a given study, but the diversity of possibilities makes it difficult to forge consensus on what to use for formal classifications of disease in regular practice.

There is as yet no standardized way of describing a particular gene expression signature derived from a specific biological source. Nor is there agreement upon what biological attribute of the signature should be used as a naming convention. Authors of the lymphochip DLBCL studies using unsupervised clustering favored using similarity to a particular stage of B-cell development as a basis for naming the signatures they identified.[8] The group that conducted supervised analysis of the DLBCL oligonucleotide array data preferred to describe the specific genes that were identified as classifiers within a given signature.[10] The ALL studies reported signatures in terms of their correlation with cytogenetic and gene fusion events.[14,15] Given the variety of descriptors appearing in the literature, the need has clearly arisen for a concise naming convention that is based upon inherent, robust properties of expression signatures.

In addition to nomenclature standardization, it is also useful to consider what might be the optimal way to structure a disease taxonomy as these new data types and descriptors emerge. In the WHO classification, histopathology remains the primary organizing principle, with genetic data inserted into relevant sub-branches of the hierarchy. This approach seems prudent: the taxonomy remains accessible to a broad spectrum of clinical practitioners, yet adds molecular criteria where appropriate. Histopathologic characterization and classification will continue to be a fundamental component of clinical practice, and provides the necessary foundation upon which molecular investigations are based.[20] Indeed, microarray studies are most appropriately focused on those areas where morphologic characterization is ambiguous.

And yet as gene expression profiling technology and nomenclature matures and robust classification criteria emerge from the data, we should explore whether treating the molecular genetic phenomena as orthogonal to histopathology might useful. This approach could prove appealing, especially if the distinguishing molecular lesions associated phenotypes are found to be common across malignancies of different histologic parentage. Expression phenotype, cytogenetic state, and immunophenotype can be correlated, but are nonetheless distinct properties of the cells in a sample. It will be challenging to develop a coherent, accessible disease taxonomy that captures and correlates, yet distinguishes these molecular attributes of a given class.

Clinical Trials and Therapy Evaluation

Classification standards can take years to evolve and achieve consensus.[21] This reality should not however prevent researchers from taking advantage of data that is available today to inform experimental design. It is now clear that several broadly classified malignancies are in fact made up of an assortment of subtypes that can be identified using expression profiling.[22] It is essential to expand the data pool for these types of analyses, so that subtypes with lower frequency in the population can be defined with the necessary statistical significance.

Grouping patients according to molecular subtype is also important for identifying the appropriate target population for a given therapy. Different molecular pathways may be involved in the different subclasses, and a given candidate therapeutic agent may have an effect on one but not another patient subgroup. This thinking is already influencing trial designs that target DLBCL patients with agents that have an impact on the NF-kB signaling pathway.[23–25]

For the above reasons, it should become common practice to include expression profiling in clinical trial protocols in these disease areas. Once collected, these data should be deposited in central repositories to enable the community to confirm the conclusions of the original publication and to perform fresh analyses as new data and methods emerge. Standards for storing and communicating the fundamental attributes of microarray

experimental data are being developed by the international Microarray Gene Expression Data (MGED) Society.[26] Once deployed, these standards will make it possible to aggregate data from different sources into larger sets for follow-on analysis. The Gene Expression Data Portal of the NCI is one such system deploying the MGED standards, and is focused on aggregating and redistributing microarray data from the kinds of cancer studies described here.[27]

CONCLUSION

Microarray methodology and expertise is not yet mature and accessible enough for the average clinical lab to deploy. Nonetheless in the controlled research setting, the approach has defined a number of specific genes that can be assayed to identify disease subclass and predict patient outcomes. It seems imperative to find near-term solutions to enable these findings to be operationally disseminated in order to provide benefit to more patients as soon as possible. One solution that has been suggested is to deploy the more manageable RT-PCR technique in routine clinical lab settings.[28] Others have shown that immunohistochemistry can be used to routinely assay for gene product expression initially found to be predictive in microarray experiments.[10] As these piece of the bench-to-bedside operations are sorted out and deployed, we will be better positioned to strike hard and perhaps even win future battles in the war on cancer.

ACKNOWLEDGEMENTS

I thank F. Hartel, N. Sioutos, and M. Heiskanen for helpful comments on the manuscript, and J. Berman for healthy skepticism.

DUALITY OF INTEREST

None declared.

Correspondence should be sent to:
Dr P Covitz, National Cancer Institute Center for Bioinformatics, 6116 Executive Boulevard, Suite 403, Rockville, MD 20852, USA
E-mail: covitzp@mail.nih.gov

1 Fragoso G, Personal communication. NCI Thesaurus, 2002; http://nciterms.nci.nih.gov.
2 Fragoso G, Personal communication. NCI Metathesaurus, 2002; http://ncimeta.nci.nih.gov.
3 Fritz A et al. (eds). *International Classification of Diseases for Oncology (ICD-O)*. World Health Organization: Geneva, 2000.
4 Jaffe ES *et al* (eds). *Pathology and Genetics of Tumours of Haematopoietic and Lymphoid Tissues*. World Health Organization: Geneva, 2001.
5 Coiffier B. *Curr Opin Oncol* 2001; **13**: 325–334.
6 Fisher RI *et al. N Engl J Med* 1993; **328**: 1002–1006.
7 Alizadeh A *et al. Cold Spring Harb Symp Quant Biol* 1999; **64**: 71–78.
8 Alizadeh AA *et al. Nature* 2000; **403**: 503–511.
9 Rosenwald A *et al. N Engl J Med* 2002; **346**: 1937–1947.
10 Shipp MA *et al. Nat Med* 2002; **8**: 68–74.
11 Slonim DK. *Nat Genet* 2002; **32** (Suppl 2): 502–508.
12 Chuaqui RF *et al. Nat Genet* 2002; **32** (Suppl): 509–514.
13 Pui CH, Evans WE *N Engl J Med* 1998; **339**: 605–615.
14 Yeoh EJ *et al. Cancer Cell* 2002; **1**: 133–143.
15 Ferrando AA *et al. Cancer Cell* 2002; **1**: 75–87.
16 Perou CM *et al. Nature* 2000; **406**: 747–752.
17 Garber ME *et al. Proc Natl Acad Sci USA* 2001; **98**: 13784–13789.
18 Lander ESJ *et al. Nature* 2001; **409**: 860–921.
19 Shoemaker DD *et al. Nature* 2001; **409**: 922–927.
20 Rosai J. *Mod Pathol* 2001; **14**: 258–260.
21 Harris NL *et al. Ann Oncol* 2000; **11** (Suppl 1): 3–10.
22 Chung CH *et al. Nat Genet* 2002; **32** (Suppl 2): 533–540.
23 Baldwin AS. *J Clin Invest* 2001; **107**: 241–246.
24 Adams J *et al. Invest New Drugs* 2000; **18**: 109–121.
25 Davis RE *et al. J Exp Med* 2001; **194**: 1861–1874.
26 Stoeckert Jr CJ *et al. Nat Genet* 2002; **32**(Suppl): 469–473.
27 Gene Expression Data Portal. National Cancer Institute, 2003; http://gedp.nci.nih.gov.
28 Rosenwald A, Staudt LM. *Semin Oncol* 2002; **29**: 258–263.